

# **EURALEX XIX**

**Congress of the  
European Association  
for Lexicography**

**Lexicography for inclusion**

**7-9 September 2021**  
**Virtual**

[www.euralex2020.gr](http://www.euralex2020.gr)

**Book of Abstracts**

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

## **The XIX EURALEX International Congress: Lexicography for inclusion**

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100  
e-edition

Publication is free of charge

# Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing

Miloš Jakubiček<sup>1,2</sup>, Vojtěch Kovář<sup>1,2</sup>, Pavel Rychlý<sup>1,2</sup>

<sup>1</sup> Lexical Computing, Brno, Czech Republic

<sup>2</sup> Masaryk University, Brno, Czech Republic

## Abstract

In this paper we report on recent findings in automatic dictionary drafting and post-editing based on two ongoing lexicographic projects, an Urdu-English-Korean dictionary and a Lao-English-Korean dictionary. We describe the basic workflow used for automatic dictionary drafting and discuss some associated methodological challenges we were facing together with solutions that we applied.

**Keywords:** Urdu, Lao, Korean, English, automatic dictionary drafting, post-editing, Sketch Engine, Lexonomy

## 1. Introduction

Contributions of natural language processing and corpus linguistics have helped lexicographers automating many parts of the dictionary building process. Recent efforts therefore focus on generating a whole dictionary draft automatically, and having it post-edited afterwards by lexicographers, roughly in the same way as translators boost their work with machine translation [3].

In this paper we illustrate this process on the example of two bilingual dictionaries: from Urdu to English and Korean and from Lao to English and Korean. These dictionaries have been drafted fully automatically and later partially post-edited. We describe the structure of the dictionaries, tools and methods used for drafting the entries and discuss management and methodological issues of the workflow we have used.

The dictionaries have been drafted from web corpora we have built and loaded into Sketch Engine [1], a corpus query system with advanced analytic functions that were used to generate the automatic draft. The post-editing phase has been carried out in Lexonomy [2], a lightweight open-source dictionary writing system.

## 2. Sketch Engine

Sketch Engine is a leading corpus management system hosting several hundreds of corpora for (as of January 2020) over 100 languages. It offers many functionalities useful for lexicographers to carry out different parts of the dictionary building, such as devising a headword list, finding good dictionary examples, generating collocation candidates or thesaurus items. All these functions have been used independently by lexicographers in many dictionary projects. In 2017 a single function combining these features has been presented under the name One-Click Dictionary:<sup>1</sup> it builds a complete dictionary draft and exports it into Lexonomy.

## 3. Lexonomy

Lexonomy is a cloud-based open-source dictionary writing and online dictionary publishing system (see more in [3]) which is highly scalable and can adapt to large dictionary projects as well as small lexicographic works such as editing and online publishing of domain-specific glossaries, wordlists or terminology resources. Lexonomy allows editing from scratch but also accepts automatically generated dictionary drafts. Lexonomy is designed to interact with Sketch Engine and its corpora in two ways: accepting generated content (“push model”) or interrogating corpora and retrieving different types of results (“pull model”).

### Push model

The push model refers to the initial dictionary draft generation. The process starts in Sketch Engine and requires that the user selects the corpus that should be used as the source data for the dictionary. Then the user decides how the dictionary headword list should be generated. Whether the dictionary headwords should be selected based on frequency using the wordlist tool or whether the headwords

---

<sup>1</sup> <https://www.youtube.com/watch?v=TaC8sTFWkqs>

should be selected from the terminology contained in the corpus using the Keywords & Terms tool. Then the user configures which parts of the dictionary entry should be generated (collocations, example sentences, synonyms, frequency information etc.). Sketch Engine then analyses the corpus and generates the required number of dictionary headwords with the required content and pushes, or exports, the automatically generated dictionary draft into Lexonomy where it is ready for further editing and for publication online.

**Pull model**

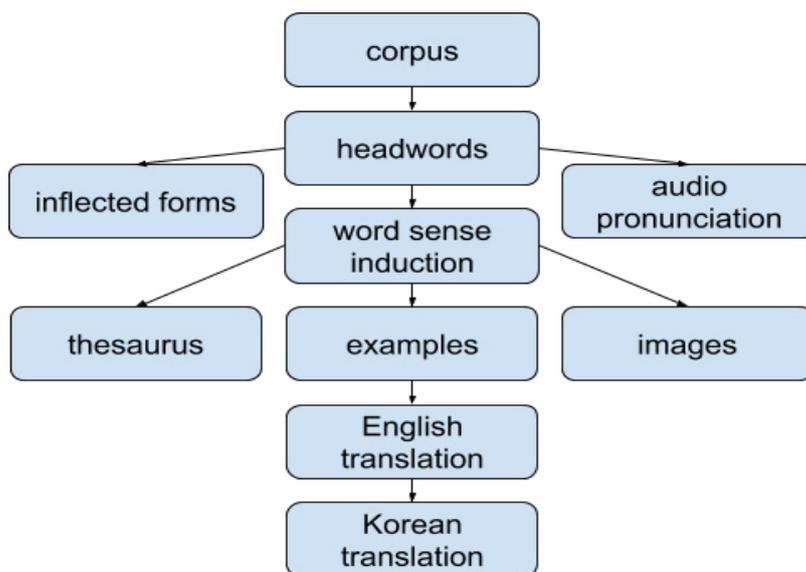
The pull model is associated with the process of post-editing the dictionary draft in Lexonomy. When the user, the dictionary editor, works with the automatically generated content, it might become necessary to check the source corpus or it might be necessary to generate additional information for the dictionary entry, for example, more collocations might be needed or different example sentences might be required. This is when the pull model comes in. Lexonomy is designed to communicate with Sketch Engine. A dictionary in Lexonomy can be linked to a specific corpus in Sketch Engine so that additional data can be pulled from the corpus if needed.

**4. From One-Click Dictionary to Million-Click Dictionary**

In the One-Click Dictionary approach, the whole dictionary draft is generated all at once. While that is useful in cases where the draft is not going to be post-edited, in the opposite case the post-editing can much more efficient if carried out step-by-step, so that errors in the automatic generation do not propagate. In this paper we argue that such a step-by-step post-editing of individual entry parts is much more time-efficient but also creates new technological and managerial issues rising from the

Figure 1: Post-editing workflow.

repetitive back-and-forth between the post-editing phase and new content generation from the corpus.



The basic workflow is described in Figure 1. Each step assumes that its parent task has been completed, clearly some of the tasks can be edited in parallel or split into a large number of batches. Key issues that we address in the paper is how to ensure data consistency and transparent backing of the underlying corpus evidence throughout the whole post-editing procedure. The reason for this is that any of the post-editing steps may result into revisions of the entry at different levels or even of the corpus material, in cases where the editor challenges the automatic corpus annotations such as part-of-speech tagging or lemmatization.

In this paper we describe our efforts on automating the management of the post-editing phase so that it would not require manual intervention between the individual post-editing tasks. We also discuss the overall efficiency of the process, based on the two dictionary projects in Urdu and Lao, each comprising 45,000 entries, out of which 15,000 have been manually post-edited.

**References**

- [1] KILGARRIFF, Adam, et al. The Sketch Engine: ten years on. *Lexicography*, 2014, 1.1: 7-36.
- [2] MĚCHURA, Michal. Introducing Lexonomy: an open-source dictionary writing and publishing system. In: *Electronic Lexicography in the 21st Century: Lexicography from Scratch*. Proceedings of the eLex 2017 conference. 2017. p. 19-21.
- [3] JAKUBICEK, Milos, et al. Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. In: *The XVIII EURALEX International Congress*. p. 65.

